Integrating Adaptive Beam-forming and Auditory Features for Robust Large Vocabulary Speech Recognition

Xie Sun, Qi (Peter) Li, Manli Zhu, and Qiru Zhou

Li Creative Technologies (LcT), Inc.
25 B Hanover Road, Suite 140, Florham Park, NJ 07932, USA
{xsun, li, manlizhu, qzhou}@licreativetech.com

Abstract

We demonstrate a system to integrate adaptive beam-forming and auditory features in order to improve speech recognition accuracy in noisy environments. A microphone array with adaptive beam-forming can utilize spatial information to improve the sound recording signal-to-noise ratio (SNR) on a focused speaker for robust speech recognition. Auditory features based on modeling the signal processing functions in the hearing system have shown to largely improve speech recognition accuracy under noisy conditions. According to our experiments, when both adaptive beam-forming and the auditory features are integrated, an absolute gain of more than 50% over the baseline on speech recognition accuracy can be achieved when the SNR is 5dB.

Index Terms: adaptive beam-forming, auditory features, speech to text (STT), SNR

1. Introduction

Compared with a mono microphone, a microphone array can be used with beam-forming algorithms to utilize spatial information during speech recognition processing. Fixed beam-forming algorithms have been shown to significantly improve robust speech recognition accuracy in [1, 2]. We introduce an adaptive beam-forming algorithm based on a 4-microphone array for speech recognition, which estimates the parameters dynamically according to the background noise, and thereby significantly reduces background noise and improves speech recognition accuracy under noisy environments.

Different feature extraction algorithms [3, 4, 5] have been used for robust speech recognition. In our recent research, an auditory based feature extraction algorithm was proposed, which has been successfully used for robust speaker identification [6]. The auditory feature extraction algorithm was further improved based on hearing system signal processing [7], which has shown to largely improve speech recognition accuracy under noisy conditions. In this demo, we present a system that integrates adaptive beam-forming and auditory features and show dramatically improvement on speech recognition performance in noisy environments.

2. System architecture

Our demo system consists of a 4-microphone array, a computational module, and a portable computer device running a speech recognition engine and application. The computational module is in charge of running the adaptive beam-forming algorithm, the feature extraction, and connecting the 4-microphone array to the portable computer via USB communication. Figure 1 illustrates our demo system.

3. Adaptive beam-forming

The structure of our adaptive beam-forming algorithm is shown in Figure 2. It consists of fixed beam-forming, a blocking matrix (BM), analysis module, adaptation module, and synthesis module. The purpose of the blocking matrix is to block the target signal and let interfering noises pass through, where those noises are fed into the adaptive filter to minimize their influence in the output. One of the key steps in adaptive beam-forming is to determine when the adaption should be applied. Because of signal leakage, the output z_i of BM may contain some weak speech signals. If the adaptation is active when speech is present, the speech will be cancelled out together with the noise. Therefore, we propose to use a module control on the adaptation. This enables adaptation according to the spectrum and energy of both noise and speech signals.

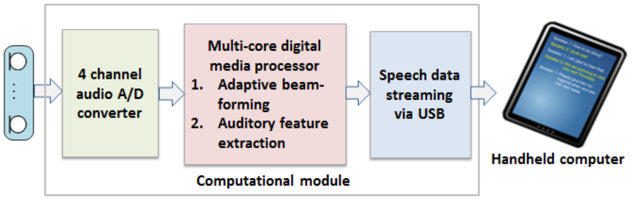
4 channel audio A/D converter
Multi-core digital media processor
1. Adaptive beam-forming
2. Auditory feature extraction
Speech data streaming via USB
Computational module
Handheld computer

Figure 1: *Demo system architecture.*

Sub-band based adaptive filtering was implemented. In the analysis step, both outputs of fixed beam-forming and BM are split into sub-bands through the analysis filter bank. In the adaptation step, the filter is adapted such that the output only contains speech signals. Finally, in the synthesis step, the sub-band speech signal is synthesized to full-band speech through the synthesis filter bank. Each sub-band adaptive filter usually has a shorter impulse response than its full-band counterpart. The step size can be adjusted individually for each sub-band, which leads to a higher convergence speed than in the case of the full-band filter.

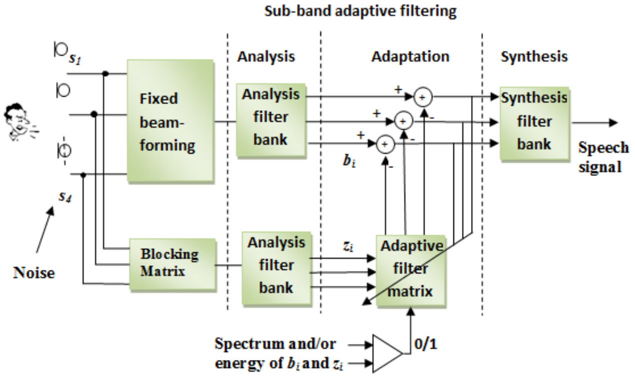

Figure 2: *Diagram of adaptive beam-forming.*

In Figure 3, we present the spectrums to show the performance of our adaptive beam-forming algorithm. The top figure is the spectrum of the original clean speech from the Wall

Street Journal (WSJ) corpus recorded by an omni-directional microphone. In addition, original clean speech was played out through a loudspeaker and recorded simultaneously using the omni-directional microphone and our 4-microphone array at the distance about 20 inches, while the white noise played out on the side through a second loudspeaker. The total signal-to-noise ratio is about 5dB. The spectrum of the omni-directional microphone recording is displayed in the middle figure and the spectrum of the microphone array recording after adaptive beam-forming is displayed in the bottom figure. Compared with the single microphone, the SNR is improved around 5dB by the microphone array with minimum effect on the speech spectrum.

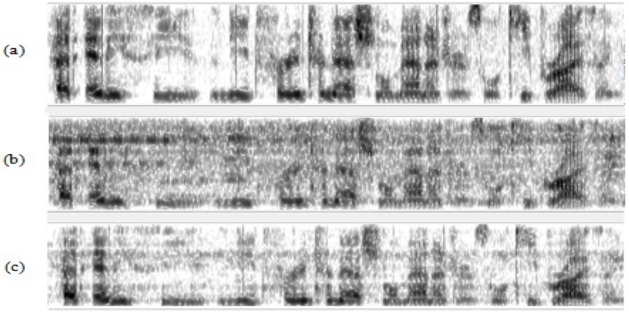

Figure 3: *Spectrums: (a) Clean speech; (b) Speech with 5dB white noise; (c) Adaptive beam-forming processed speech.*

4. Auditory features

A new auditory features from [7] were successfully developed recenty for robust speach recognition., which can significantly improve speech recognition performance under noisy environments. An illustrative block diagram of the hearing model used to extract the auditory features is shown in Figure 4 [7]. It consists of the following modules: auditory transform, energy normalization, meddis hair cell model, equal-loudness function, windowing, loudness nonlinearity, and discrete cosine transform (DCT). Except for DCT, all the modules are for modeling the signal processing functions of the human hearing system. We name the new features from the model as auditory feature cepstral coefficients (AFCC). For a detailed explanation of each module, please refer to [7].

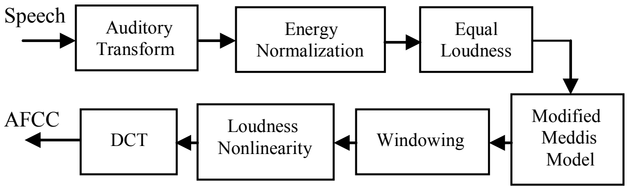

Figure 4. *Diagram of auditory feature extraction [7]* .

Table 1 shows a comparison of the AFCC features with MFCC (Mel-frequency cepstral coefficient), PLP (perceptual linear prediction), and RASTA-PLP features in word accuracy on a speaker independent ASR task at different SNR levels. The experiments were implemented based on the WSJ corpus. The training data is clean speech from WSJ-SI84 [8] and the test data is noisy speech based on the standard Nov 92 test set [8]. Cepstral mean normalization was used on all types of features. More detail about these experiments can be found in [7]. Overall, we see that the proposed AFCC features outperform other features in this large-vocabulary speech recognition task under mismatched and noisy conditions.

Table 1. Performance comparison of features in word accuracy (%) on Test Set with Added White Noise

Testing SNR	5dB	10dB	15dB	20dB	Clean
AFCC	36.02	65.53	78.95	85.34	89.20
MFCC	10.61	31.76	64.99	82.22	90.88
PLP	12.91	35.10	66.11	82.70	91.01
RASTA-PLP	12.89	36.33	66.02	80.40	87.46%

5. Demo designs

We demonstrate our speech recognition system under noisy environments using three demos. In all three demos, clean speech is used to train acoustic models and white noise is added to test speech with different SNRs such as 5dB, 10dB and 15dB. In the first demo, we compare speech recognition performance between speech processed with and without adaptive beam-forming and all other conditions are the same for these two cases. In our experiments, for instance, adaptive beam-forming improved the SNR from 5dB to 10dB and the absolute gain for speech recognition accuracy improved more than 20%. In the second demo, we compare speech recognition accuracy between AFCC and MFCC features. As shown in Table 1, for example, the AFCC features show more than 25% absolute improvement over the MFCC features when the SNR is 5dB. In the third demo, we integrate adaptive beam-forming and auditory features to further improve speech recognition accuracy under noisy environments. The auditory features are extracted based on the adaptive beam-forming processed speech. According to our experimental results, the integration led to an absolute gain of more than 50% improvement over a MFCC based baseline system in robust speech recognition when SNR level is 5dB.

6. Conclusion

We have proposed to integrate adaptive beam-forming and auditory features for speech recognition, which can largely improve speech recognition accuracy under noisy environments. Our goal is to contribute these new technologies as the new front-end of speech recognition under noisy environments.

7. References

[1] Q. Li, M. Zhu, and W. Li, “A portable usb-based microphone array device for robust speech recognition,” in *ICASSP 2009*, 2009.

[2] A. Stolcke, “Making the most from multiple microphones in meeting recognition,” in *ICASSP 2011*, 2011.

[3] D. Dimitriadis, E. Bocchieri and D. Caseiro, "An alternative front-end for the AT&T Watson LV-CSR system," in *ICASSP 2011*, 2011.

[4] C. Kim and R. M. Stern. “Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction,” in *INTERSPEECH 2009*, pp. 28-31, 2009.

[5] H. Hermansky and N. Morgan, “RASTA processing of speech,” in *IEEE Trans. on Speech and Audio Proc.*, vol. 2, pp. 578–589, Oct. 1994.

[6] Q. Li and X. Sun, “Feature extraction based on hearing system signal processing for robust large vocabulary speech recognition,” in *INTERSPEECH 2012*, 2012 (submitted).

[7] P. C. Woodland, J. J. Odell, V. Valtech, and S. J. Young, “Large vocabulary continuous speech recognition using HTK,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 125–128, 1994.